

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Scalable Factorization Model to Discover Implicit
and Explicit Similarities Across Domains**

by

Duc Minh Quan Do

A THESIS SUBMITTED FOR THE DEGREE OF

Doctor of Philosophy

Sydney, Australia

2018

UNIVERSITY OF TECHNOLOGY SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled “**Scalable Factorization Model to Discover Implicit and Explicit Similarities Across Domains**” by **Duc Minh Quan Do** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Date:

Principal Supervisor: _____

Dr. Wei Liu

Certificate of Original Authorship

I, Duc Minh Quan Do declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) scholarship.

Date: 15/09/2018

Signature of Author:	Production Note:
	Signature removed prior to publication.

Acknowledgements

I am especially indebted to Dr. Wei Liu, who have provided continuous support, advice and invaluable comments to pursue my research goals. As my principal supervisor, he has guided me more than I could ever give him credit for here. Many thanks are also due to my co-supervisor, Dr. Fang Chen, for many useful discussions with her.

I am grateful to all I have had the pleasure to discuss with. Each of the members of my Candidature Assessment Committee has provided me a great deal of professional feedback about scientific research. This work would not have been possible without the financial support of the Commonwealth Scientific and Industrial Research Organisation Scholarship (formerly National ICT Australia Scholarship) and the UTS - International Research Scholarship (IRS).

Nobody has been more important to me in the pursuit of this thesis than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue and wherever I am. They are the ultimate role models. Most importantly, I am grateful to my loving and supportive wife, Yen, and my wonderful daughter, Ellen, for constant inspiration, patience, and faith.

For Ellen

My love for you will last forever.

Contents

Certificate	iii
Acknowledgments	iv
Dedication	v
List of Figures	xi
List of Tables	xiv
List of Publications	xv
Abbreviation	xvi
Notation	xvii
Abstract	xix
1 Introduction	1
1.1 The research problem	3
1.1.1 The improper sharing of explicit similarities among coupled datasets across domains reduces recommendation accuracy . .	3
1.1.2 Coupled datasets across domains also share implicit similarities that provide other insights into their relationships	5
1.1.3 Joint analysis of heterogeneous datasets is costly	7
1.2 Thesis	8
1.3 Background	11
1.4 Knowledge contributions	12

1.5	Research Methods	13
1.5.1	A new objective function to enable each dataset to have its own discriminative factor on the coupled mode, capturing the actual explicit similarities across domains	14
1.5.2	A novel algorithm to discover implicit similarities in non-coupled mode and align them across domains	16
1.5.3	A matrix factorization-based model to utilize both explicit and implicit similarities for cross-domain recommendation accuracy improvement	17
1.5.4	A scalable factorization model based on the Spark framework to scale up the factorization process to the number of tensors, tensor modes, tensor dimensions and billions of observations	18
1.6	Significance	19
1.7	Thesis organization	21
2	Literature Review and Background	24
2.1	Data format	25
2.1.1	Rating matrix (utility matrix)	25
2.1.2	Tensor	26
2.1.3	Coupled datasets	27
2.2	Recommendation Systems	28
2.2.1	Matrix Factorization	29
2.2.2	Matrix Tri-Factorization	30
2.2.3	Tensor Factorization	30
2.3	Cross-domain Recommendation Systems	32
2.3.1	Collective Matrix Factorization	32

2.3.2	Coupled Matrix Tensor Factorization	33
2.3.3	CodeBook Transfer	35
2.3.4	Cluster-Level Latent Factor Model	36
2.4	Factorization Methodologies	36
2.5	Distributed Factorization	37
2.6	Deep learning based recommendation systems	39
2.7	Research gaps	40
3	Explicit Similarity Discovery	44
3.1	Introduction	44
3.2	ASTEN: the proposed Accurate Coupled Tensor Factorization model .	46
3.3	Optimization	47
3.4	Performance Evaluation	50
3.4.1	Data used in our experiments	50
3.4.2	Performance metric	52
3.4.3	Results	54
3.5	Contribution and Summary	56
4	Implicit Similarity Discovery	58
4.1	Introduction	58
4.2	HISF: the proposed Hidden Implicit Similarities Factorization Model .	61
4.2.1	Sharing common and preserving domain-specific coupled latent variables to utilize explicit similaritites	62
4.2.2	Aligning implicit similarities in non-coupled latent clusters across domains	62
4.2.3	Optimization	71

4.3	Extension to three or more matrices	76
4.4	Experiments and Analysis	78
4.4.1	Data for the experiments	79
4.4.2	Experimental settings	80
4.4.3	Empirical results	81
4.5	Contributions and Summary	88
5	Scalable Multimodal Factorization	91
5.1	Introduction	91
5.2	SMF: the proposed Scalable Multimodal Factorization	93
5.2.1	SMF on Apache Spark	96
5.2.2	Scaling up to K tensors	102
5.3	Performance Evaluation	103
5.3.1	Scalability	105
5.3.2	Convergence Speed	106
5.3.3	Accuracy	109
5.3.4	Optimization	110
5.4	Contribution and Summary	111
6	Conclusion	114
6.1	Research questions and contributions	115
6.2	Future research directions	118
6.2.1	Investigating explicit and implicit similarities in imbalanced datasets	118
6.2.2	Extending the use of explicit and implicit similarities to high dimensional tensors	119

6.2.3	Extending the proposed factorization model to handle online ratings	119
6.2.4	Investigating the use of explicit and implicit similarities in Factorization Machines	119
6.3	Conclusion	120

List of Figures

1.1	An example of implicit similarities	5
1.2	The research questions and their correspondent contributions	13
2.1	An example of a movie rating matrix	26
2.2	An example of a tensor	27
2.3	An example of coupled rating matrices from Netflix and MovieLens websites	27
2.4	An example of a coupled matrix tensor from MovieLens website . . .	28
2.5	CANDECOMP/ PARAFAC (CP) decomposition	31
2.6	Joint analysis of a coupled matrix tensor	34
2.7	Distributed factorization algorithms	38
2.8	Multi-view deep neural network for cross-domain recommendation two datasets where they have the same users. In this case, users of both datasets share the same features of the left-most network.	40
3.1	Mean squared errors of test cases with synthetic data	53
3.2	Mean squared error of factorizing the MovieLens dataset	54
3.3	Mean squared error of factorizing Yahoo! Music dataset	55

4.1	The proposed factorization model to discover and share implicit similarities across domains	63
4.2	Matrix factorization of $\mathbf{X}^{(1)}$ as a clustering method	64
4.3	Matrix factorization of $\mathbf{X}^{(2)}$ as a clustering method	65
4.4	Possible cases for matching user clusters of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$	66
4.5	An illustration of how centroid of a cluster is computed	68
4.6	Generated ratings of two domains $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$	69
4.7	An illustration of how well the proposed cluster alignment method works	70
4.8	Tested mean RMSEs of ABS NSW and ABS VIC datasets under different values of the common row parameter (c) in the coupled factor of HISF with rank $r = 11$	84
4.9	Tested mean RMSEs of ABS NSW and BOCSAR Crime datasets under different values of the common row parameter (c) in the coupled factor of HISF with rank $r = 11$	85
4.10	Tested mean RMSEs of Amazon dataset under different values of the common row parameter (c) in the coupled factor of HISF-N with rank $r = 11$	88
5.1	Tensor slices for updating each row of the factors when a mode-3 tensor is coupled with a matrix in their first modes	95
5.2	An example of how to divide coupled matrix and tensor into non-overlapping blocks	97
5.3	Observation scalability	106
5.4	Machine scalability with 100M synthetic dataset	107
5.5	Factorization speed with MovieLens	107
5.6	Factorization speed with Netflix	108

5.7	Factorization speed with Yahoo! Music	108
5.8	Coupled factorization speed with MovieLens	108
5.9	Coupled factorization speed with Yahoo! Music	109
5.10	Benchmark of different optimization methods	110

List of Tables

1	Symbols and their descriptions	xviii
1.1	Comparison of existing algorithms for recommendation	9
3.1	Ground truth distributions of the factor matrices in the synthetic data	51
4.1	Characteristics of ABS census data on New South Wales and Victoria states	79
4.2	Characteristics of Amazon datasets on books, movies and electronics	80
4.3	Mean and standard deviation of tested RMSE on ABS New South Wales and Victoria data with different algorithms	81
4.4	Mean and standard deviation of tested RMSE on ABS NSW demography and BOCSAR NSW crime data with different algorithms	84
4.5	Mean and standard deviation of tested RMSE on Amazon book, movie and electronics data with different algorithms	86
5.1	Data for experiments	105
5.2	Accuracy of each algorithm on the real-world datasets	109
5.3	Accuracy of predicting missing entries on real-world datasets with different optimizers	111

List of Publications

Below is the list of journal and conference papers associated with my Ph.D. research:

1. **Quan Do**, Wei Liu, Fan Jin and Dacheng Tao, “Unveiling Hidden Implicit Similarities for Cross-Domain Recommendation,” *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (Under review).
2. **Quan Do** and Wei Liu, “Scalable Multimodal Factorization for Learning from Very Big Data,” in *Multimodal Analytics for Next-Generation Big Data Technologies and Applications*, Springer (To appear).
3. **Quan Do**, Wei Liu and Fang Chen, “Discovering both Explicit and Implicit Similarities for Cross-Domain Recommendation,” in *Proceedings of the 2017 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 618-630, May 23-26, 2017.
4. **Quan Do** and Wei Liu, “ASTen: an Accurate and Scalable Approach to Coupled Tensor Factorization,” in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 99-106, Jul. 24-29, 2016.

Abbreviation

ALS Alternating Least Squares. 14

CBT Code Book Transfer. 6, 18

CF Collaborative Filtering. x, 1, 11

CLFM Cluster-Level Latent Factor Model. 6, 8, 18

CMF Collective Matrix Factorization. 2, 8, 13, 17

CMTF Coupled Matrix Tensor Factorization. 5, 6, 8, 11, 13–17

GD Gradient Descent. 14

GPU Graphics Processing Unit. 15

MF Matrix Factorization. 11, 12

NCG Nonlinear Conjugate Gradient. 14

NSW New South Wales. 1, 4

RMSE Root Means Squared Error. 6

TF Tensor Factorization. 11, 12, 14, 15

Nomenclature and Notation

A rating matrix from n users for m items is denoted by a boldface capital, e.g., \mathbf{X} . Each row of the matrix is a vector of a user's ratings for all items while each column is a vector of ratings from all users for a specific item. Vectors are denoted by boldface lowercases, i.e., \mathbf{u} . A boldface capital and lowercase with indices in their subscript are respectively used for an entry of a matrix and a vector. Table 1 lists all other symbols we thoroughly use in this thesis.

Table 1 : Symbols and their descriptions

Symbol	Description
$\mathbf{X}^{(i)}$	Rating matrix from i -th dataset
$\mathbf{U}^{(i)}$	The first dimension factor of $\mathbf{X}^{(i)}$
$\mathbf{V}^{(0)}$	Common parts of the coupled factors
$\mathbf{V}^{(i)}$	Domain-specific parts of the coupled factor of $\mathbf{X}^{(i)}$
$\mathbf{S}^{(i)}$	Weighting factor of $\mathbf{X}^{(i)}$
\mathbf{A}^T	Transpose of \mathbf{A}
\mathbf{A}^\dagger	Moore-Penrose pseudo inverse of \mathbf{A}
\mathbf{I}	The identity matrix
$\ \mathbf{A}\ $	Frobenius norm
n, m, p	Dimension length
c	Number of common clusters in coupled factors
r	Rank of decomposition
$\Omega_{\mathbf{X}}$	Number of observations in \mathbf{X}
$\frac{\partial}{\partial \mathbf{x}}$	Partial derivative with respect to \mathbf{x}
\mathcal{L}	Loss function
λ	Regularization parameter
\times	Multiplication
$x, \mathbf{x}, \mathbf{X}, \mathfrak{X}$	A scalar, a vector, a matrix and a tensor
N	Mode of a tensor
M	Number of machine
K	Number of tensor
T	Number of iteration
$I_1 \times I_2 \times \cdots \times I_N$	Dimension of N -mode tensor \mathfrak{X}
$ \Omega , \mathfrak{X}_{i_1, i_2, \dots, i_N}$	Observed data size of \mathfrak{X} and its entries
$\mathfrak{X}^{(n)}$	Mode n^{th} of \mathfrak{X}
$\mathfrak{X}_{i_n}^{(n)}$	Slice i_n of $\mathfrak{X}^{(n)}$ - all entries $\mathfrak{X}_{*, \dots, *, i_n, *, \dots, *}^{(n)}$
$\mathbf{U}^{(n)}$	n^{th} mode factor of \mathfrak{X}
$\mathbf{u}_{i_n}^{(n)}$	i_n^{th} row of factor $\mathbf{U}^{(n)}$
$\mathbf{V}^{(2)}$	2^{nd} mode factor of \mathbf{Y}
$\mathbf{v}_{j_2}^{(2)}$	j_2^{th} row of factor $\mathbf{V}^{(2)}$ - all entries $\mathbf{V}_{*, j_2}^{(2)}$
$\mathbf{U}1, \mathbf{U}2, \dots, \mathbf{U}K$	Factors of tensor $\mathfrak{X}1, \mathfrak{X}2, \dots, \mathfrak{X}K$
$I_1 \times I_2 \times \cdots \times I_{N_K}$	Dimension of N_K -mode tensor $\mathfrak{X}K$
$ \Omega _K, \mathfrak{X}K_{i_1, i_2, \dots, i_{N_K}}$	Observed data size of $\mathfrak{X}K$ and its entries

Abstract

E-commerce businesses increasingly depend on recommendation systems to introduce personalized services and products to their target customers. Achieving accurate recommendations requires a sufficient understanding of user preferences and item characteristics. Given the current innovations on the Web, coupled datasets are abundantly available across domains. An analysis of these datasets can provide a broader knowledge to understand the underlying relationship between users and items. This thorough understanding results in more collaborative filtering power and leads to a higher recommendation accuracy.

However, how to effectively use this knowledge for recommendation is still a challenging problem. In this research, we propose to exploit both explicit and implicit similarities extracted from latent factors across domains with matrix tri-factorization. On the coupled dimensions, common parts of the coupled factors across domains are shared among them. At the same time, their domain-specific parts are preserved. We show that such a configuration of both common and domain-specific parts benefits cross-domain recommendations significantly. Moreover, on the non-coupled dimensions, the middle factor of the tri-factorization is proposed to use to match the closely related clusters across datasets and align the matched ones to transfer cross-domain implicit similarities, further improving the recommendation.

Furthermore, when dealing with data coupled from different sources, the scalability of the analytical method is another significant concern. We design a distributed factorization model that can scale up as the observed data across domains increases. Our data parallelism, based on Apache Spark, enables the model to have the smallest communication cost. Also, the model is equipped with an optimized solver that converges faster. We demonstrate that these key features stabilize our model's per-

formance when the data grows.

Validated on real-world datasets, our developed model outperforms the existing algorithms regarding recommendation accuracy and scalability. These empirical results illustrate the potential of our research in exploiting both explicit and implicit similarities across domains for improving recommendation performance.